Agrégation optimale sous contrainte de contiguïté

Marc CHRISTINE
Michel ISNARD
Insee

Plan de la présentation

- 1. Introduction
- 2. Deux exemples introductifs
- 3. Description rapide de la méthode
- 4. Applications à des exemples simples
- 5. Application à une distance non euclidienne
- 6. Conclusion

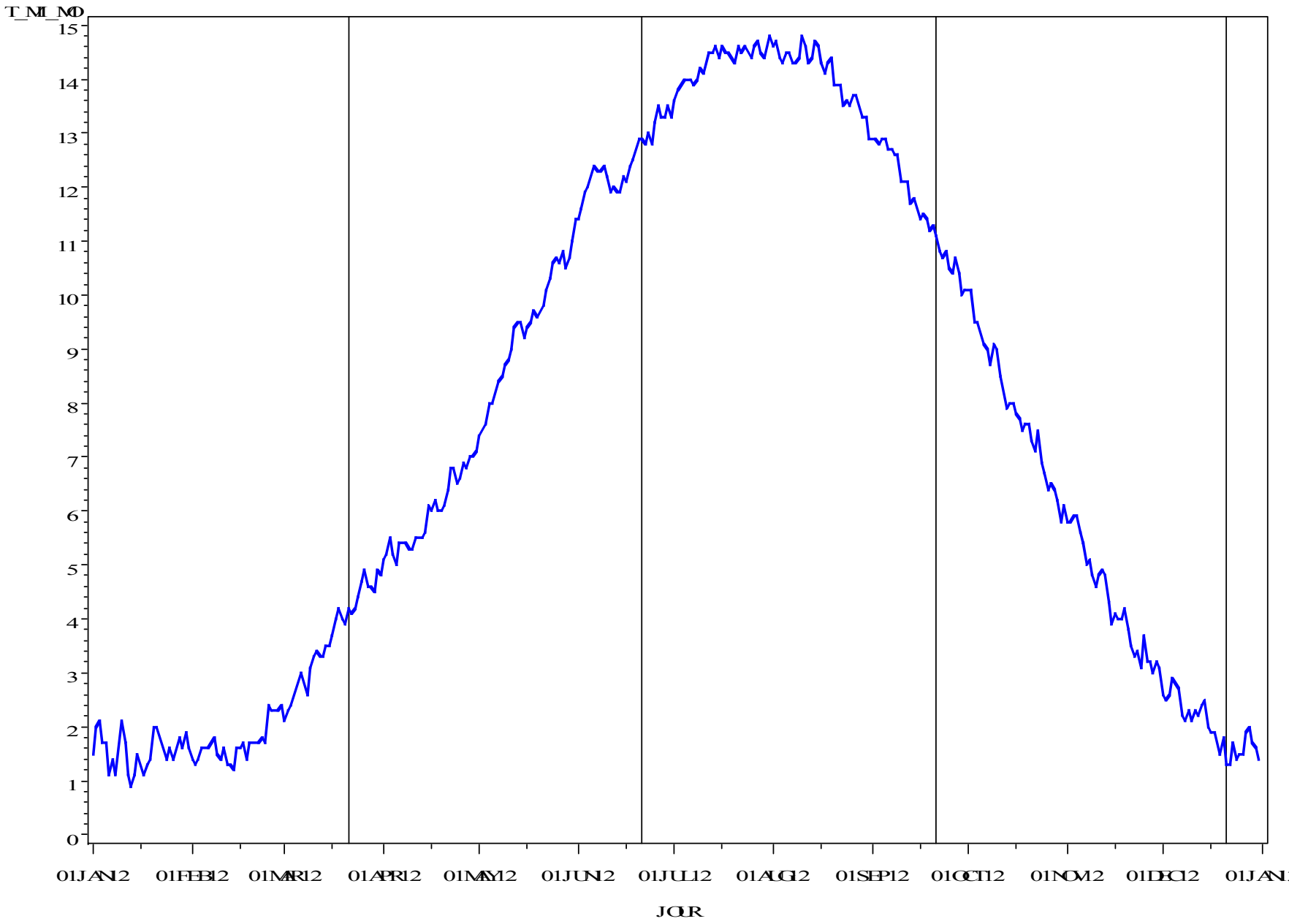
1. Que veut-on faire ?

- A partir
  - d'une population d'unités statistiques pondérées
  - entre lesquelles a été définie une distance (euclidienne ou pas) …
  - … et une relation de contiguïté

- Partager la population en K (nombre fixé par l'utilisateur) groupes connexes, vérifiant des contraintes uniformes de taille (également fixées par l'utilisateur) et minimisant ou maximisant l'inertie intra-groupe.

2. Deux exemples introductifs

Températures minimales moyennes à Paris depuis 1873



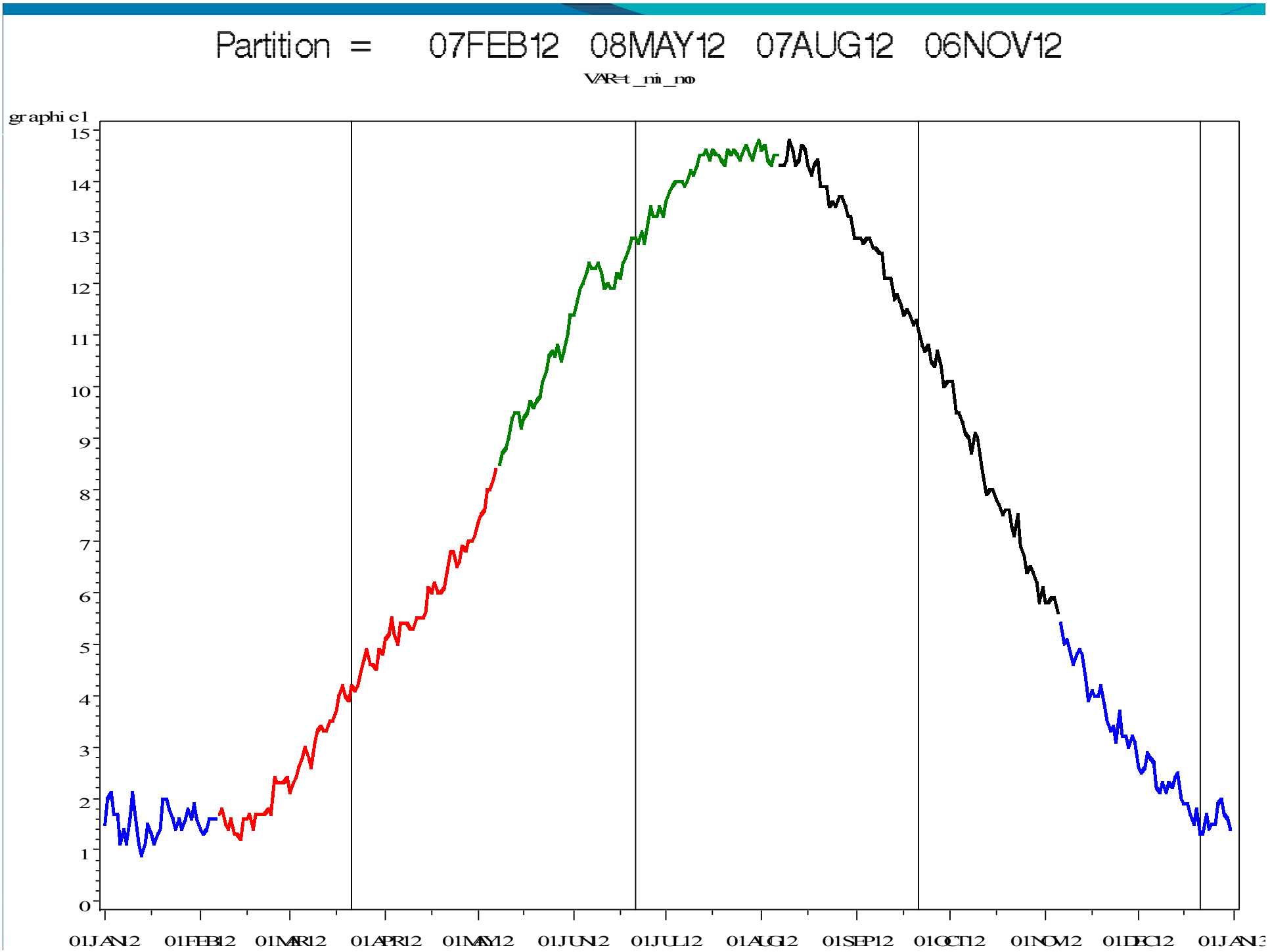
Données recueillies sur le site www.meteo-paris.com

Comment recréer des saisons ?

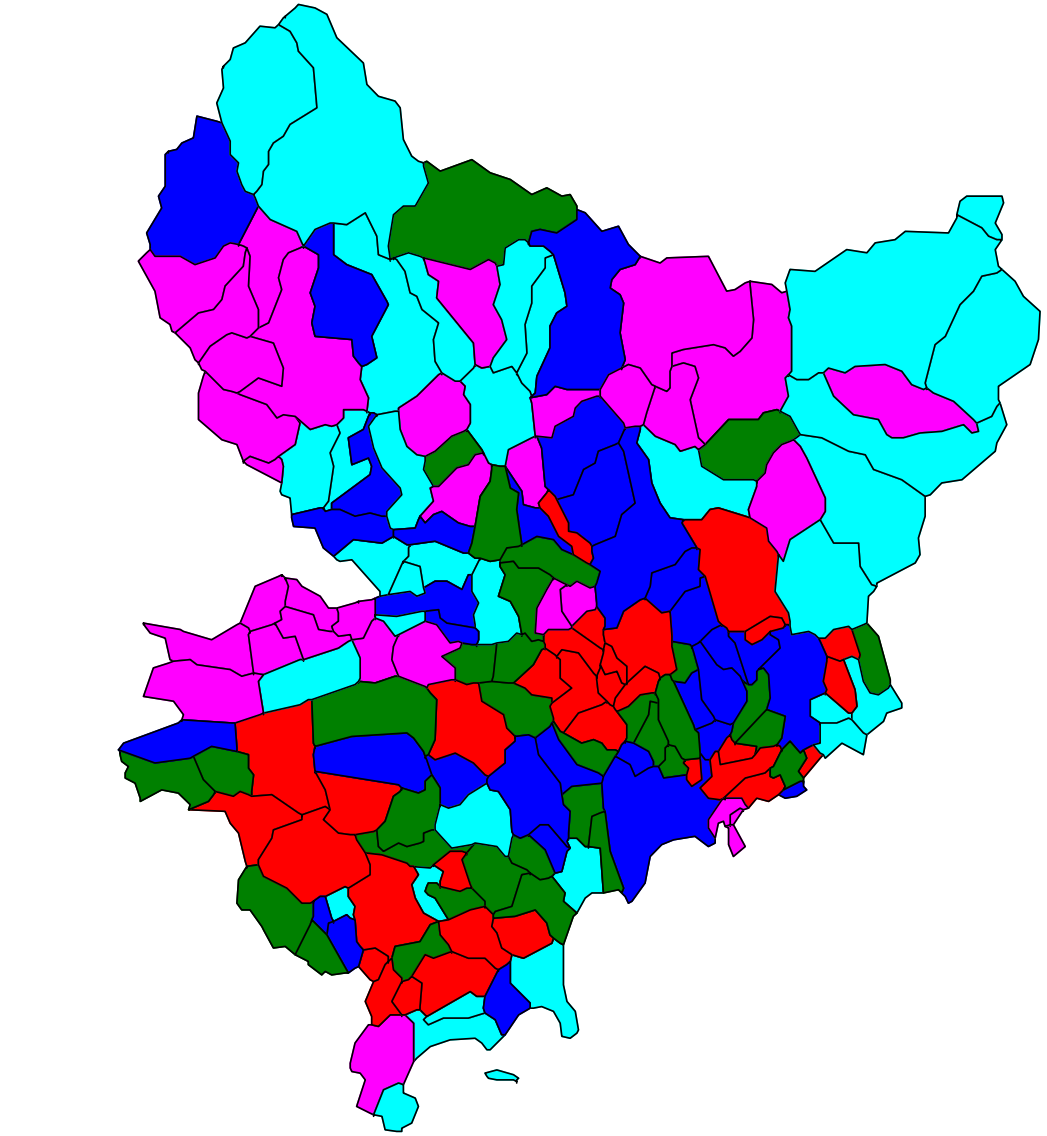
- Trouver toutes les partitions en plages de longueur comprises en 91 et 93 jours
  - 1er janvier et 31 décembre se touchant
- Calculer la variance intra-groupe de chacune de ces partitions (pour la variable « température »)
- Trouver la partition ayant la plus petite variance intra-groupe
  - Saisons les plus homogènes possibles.

Comment recréer des saisons ?

- 920 partitions possibles


Partition = 07FEB12 08MAY12 07AUG12 06NOV12
VAR=t_mi_no
graphic1
15
14
13
12
11
10
9
8
7
6
5
4
3
2
1
0
01JAN12
01FEB12
01MAR12
01APR12
01MAY12
01JUN12
01JUL12
01AUG12
01SEP12
01OCT12
01NOV12
01DEC12
01JAN13

Age moyen dans les communes des Alpes Maritimes



Source - Insee - RP

Nombre de partitions à tester

- 10^{111} partitions en 5 classes des 163 communes des Alpes Maritimes
  - Et … $0{,}000\ldots\varepsilon$ % partitions connexes
- Nécessité de trouver une autre méthode…
- … et de fournir un algorithme et un programme permettant d'atteindre des solutions

3. Description rapide de la méthode

- Une phase d'agrégation proche d'une classification ascendante hiérarchique
- Une seconde phase d'échange entre les groupes créés précédemment afin de respecter les critères de taille et d'optimiser l'inertie intra-groupe

Les contraintes initiales

- Le programme doit fournir des groupes connexes respectant au mieux les contraintes données par l'utilisateur.

- L'utilisateur désirant utiliser cette méthode doit :
  - Fournir une liste d'unités statistiques pondérées
  - … et une distance entre deux quelconques de ces unités (construite à partir d'une variable d'intérêt)
  - Indiquer les unités contiguës
  - Indiquer le nombre de groupes qu'il souhaite créer et les contraintes de taille (uniformes) qu'il souhaite voir respecter
  - Indiquer s'il souhaite maximiser ou minimiser l'inertie (=variance) intra-groupe.

La première phase

- Bâtie sur le même algorithme qu'une CAH
- Mais on limite l'agrégation aux groupes *contigus*, c'est-à-dire aux groupes dont au moins une unité est contiguë à une unité de l'autre groupe.
- A chaque étape, on choisit les groupes à agréger : ceux dont l'agrégation minimise ou maximise la variation d'inertie intra-groupe

- **Les groupes formés sont des groupes connexes, mais ne respectant pas forcément les contraintes de taille**

La seconde phase

- A partir des groupes créés lors de la première phase,
  - On améliore le respect des contraintes de taille (et d'optimisation de l'inertie)
  - En échangeant des unités de groupe à groupe ou en transférant des unités d'un groupe à un autre

- L'expérience montre que les contraintes de taille sont respectées, dès lors qu'elles ne sont pas trop sévères.

La méthode en résumé

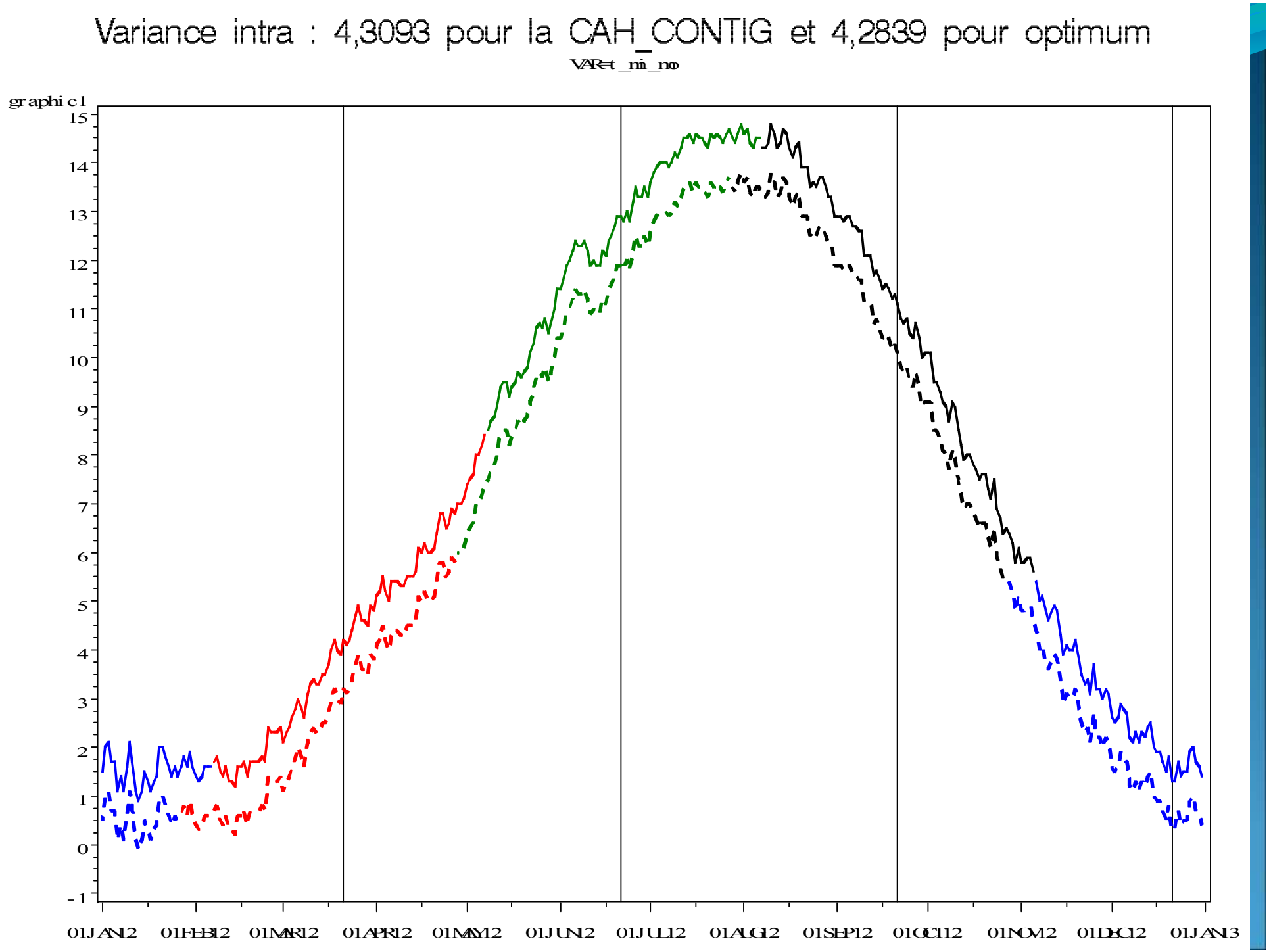
- Tous les groupes créés sont des groupes connexes
- Les contraintes de taille sont très fréquemment vérifiées
- On n'est pas assuré de l'optimalité de l'inertie

4. Retour sur les saisons

- Maximisation de l'inertie intra-saison (= variance de la température) avec des saisons de longueur comprise en 91 et 93 jours

- Minimisation de l'inertie intra-saison avec les mêmes contraintes

Variance intra : 4,3093 pour la CAH_CONTIG et 4,2839 pour optimum



Retour sur les saisons

- Limites différentes :
  - 27 JAN, 27 AVR, 27 JUL et 26 OCT pour CAH_CONTIG
  - 07 FEV, 08 MAY, 07 AUG et 06 NOV pour l'optimum

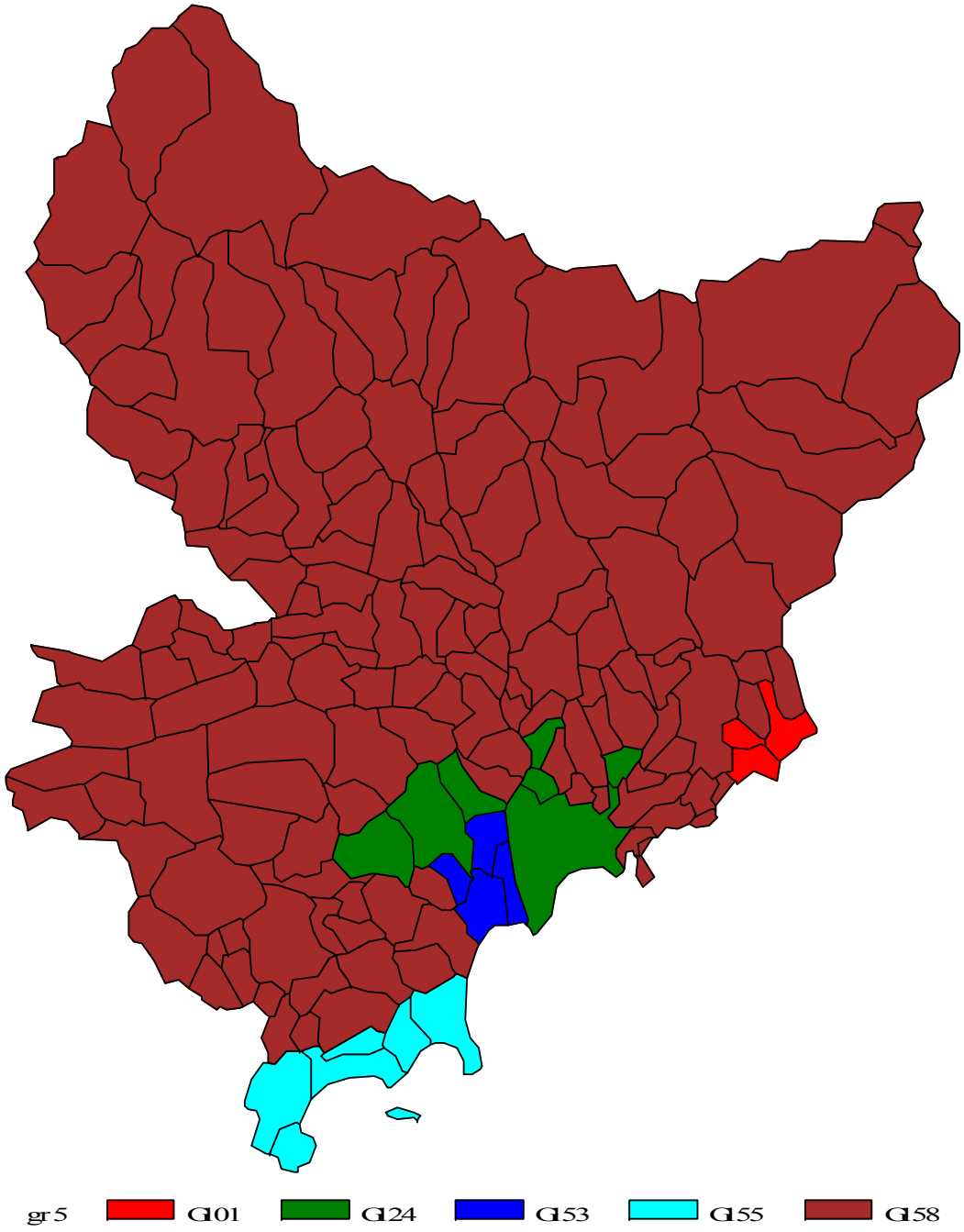
- Variances intra-groupe
  - 4,3093 pour la CAH CONTIG
  - 4,2839 pour l'optimum

- NON OPTIMAL … (mais écart faible)

Un exemple sur les Alpes-Maritimes

- Variable : Age moyen par commune (RP 2008)
  - Minimisation de la variance intra-groupe : regroupement en 5 groupes connexes les plus homogènes possibles,
  - Contraintes de taille : chaque groupe doit comprendre entre 15 et 35% de la population

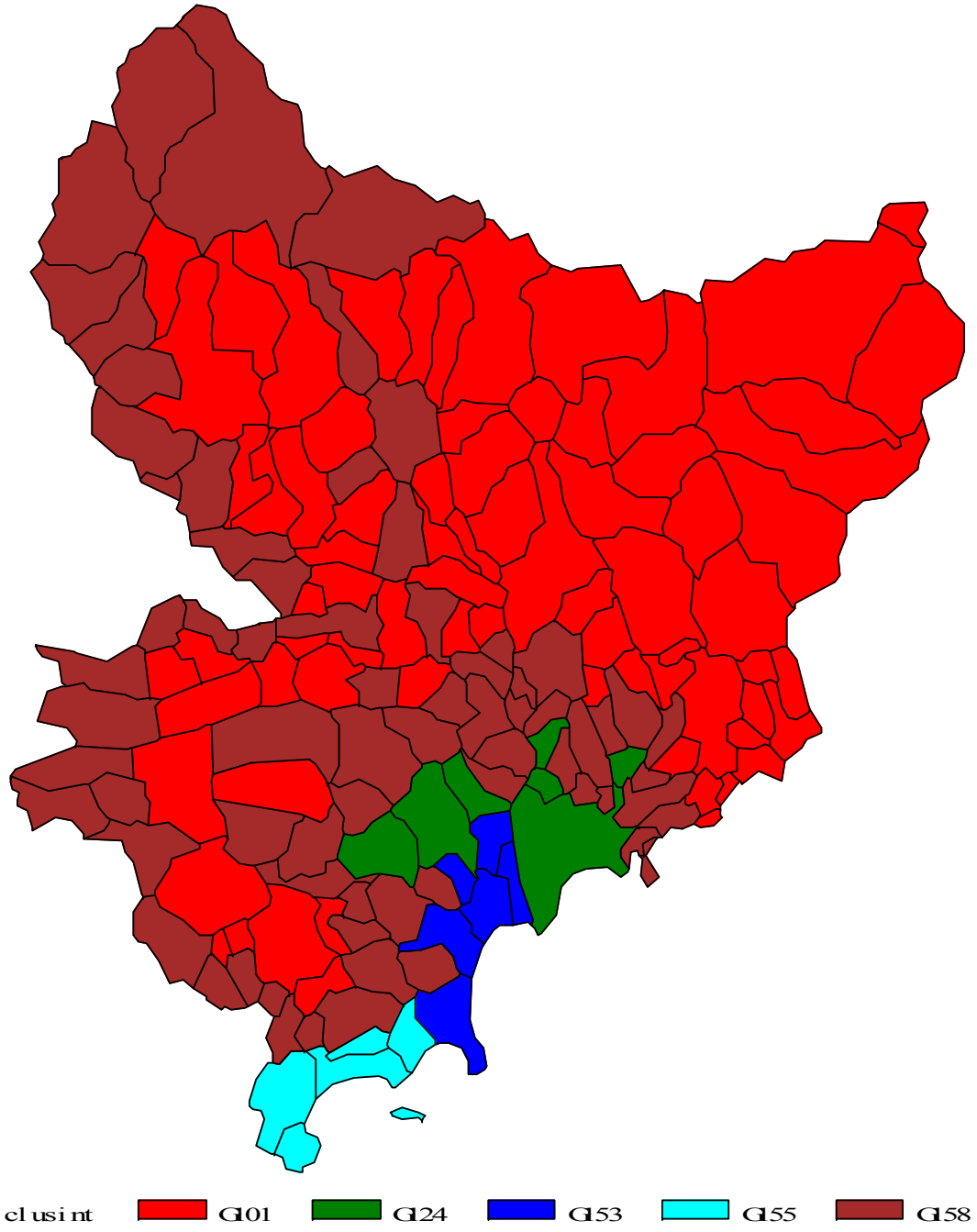
Phase 1 – Minimisation



Résultats de la première phase

Groupes	Taille	Age Moyen	Variance intra
G101	4,0%	45,4	0,002
G124	34,8%	43,1	0,013
G153	8,2%	43,2	0,126
G155	22,4%	45,6	0,309
G158	30,5%	41,0	2,752
		43,1	**3,202**

Phase 2 – Minimisation



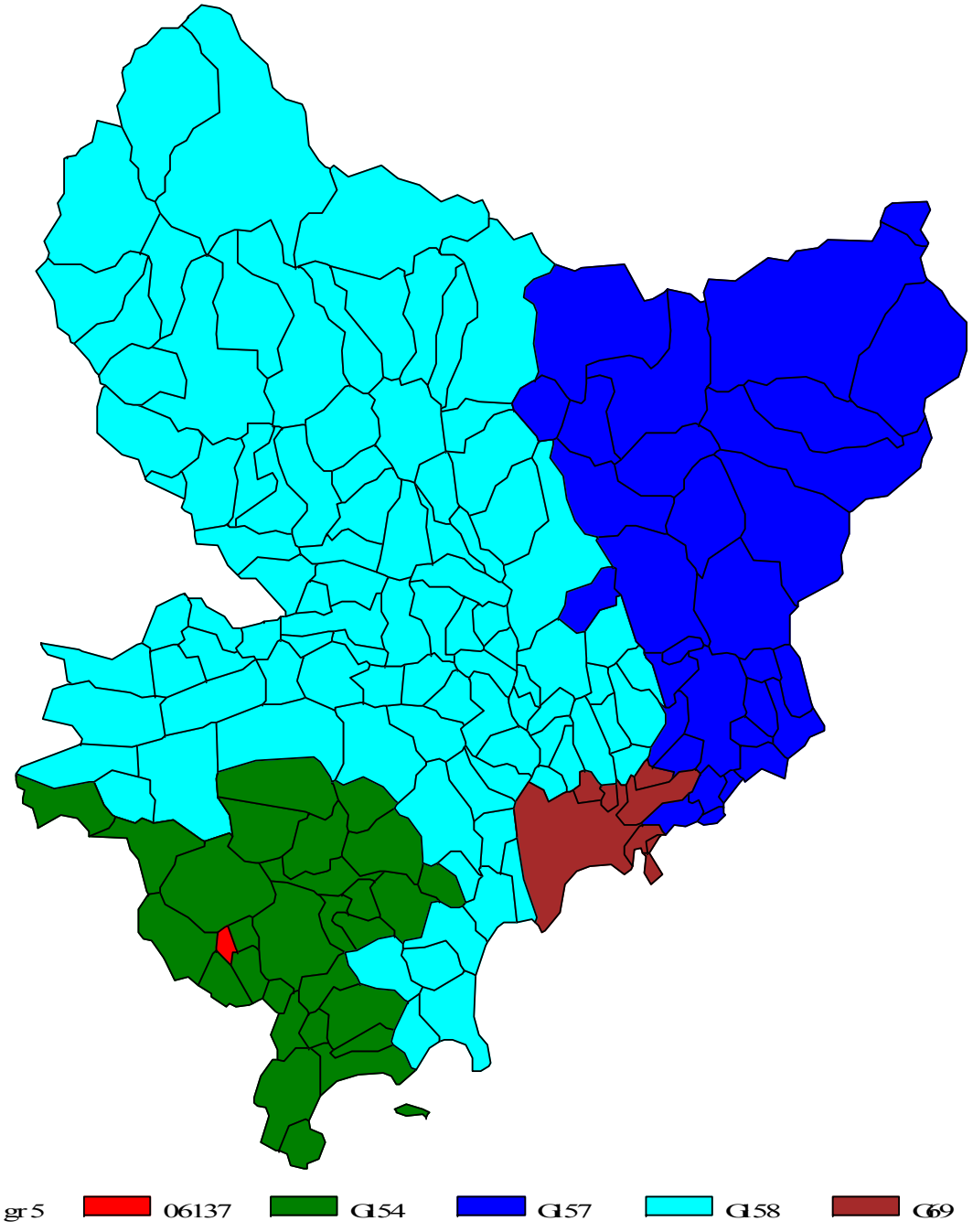
Résultats de la seconde phase

Groupes	Taille	Age moyen	Variance intra
G101	16,1%	42,5	1,525
G124	34,8%	43,1	0,013
G153	16,7%	43,7	0,378
G155	15,3%	45,9	0,267
G158	17,1%	40,8	1,65
		43,1	**3,830**

Un second exemple sur les Alpes-Maritimes

- Variable : Age moyen par commune (RP 2008)
  - **Maximisation** de la variance intra-groupe : regroupement en 5 groupes connexes les plus proches possibles de la population totale,
  - Contraintes de taille : chaque groupe doit comprendre entre 15 et 35% de la population
- Inertie totale = 6,088

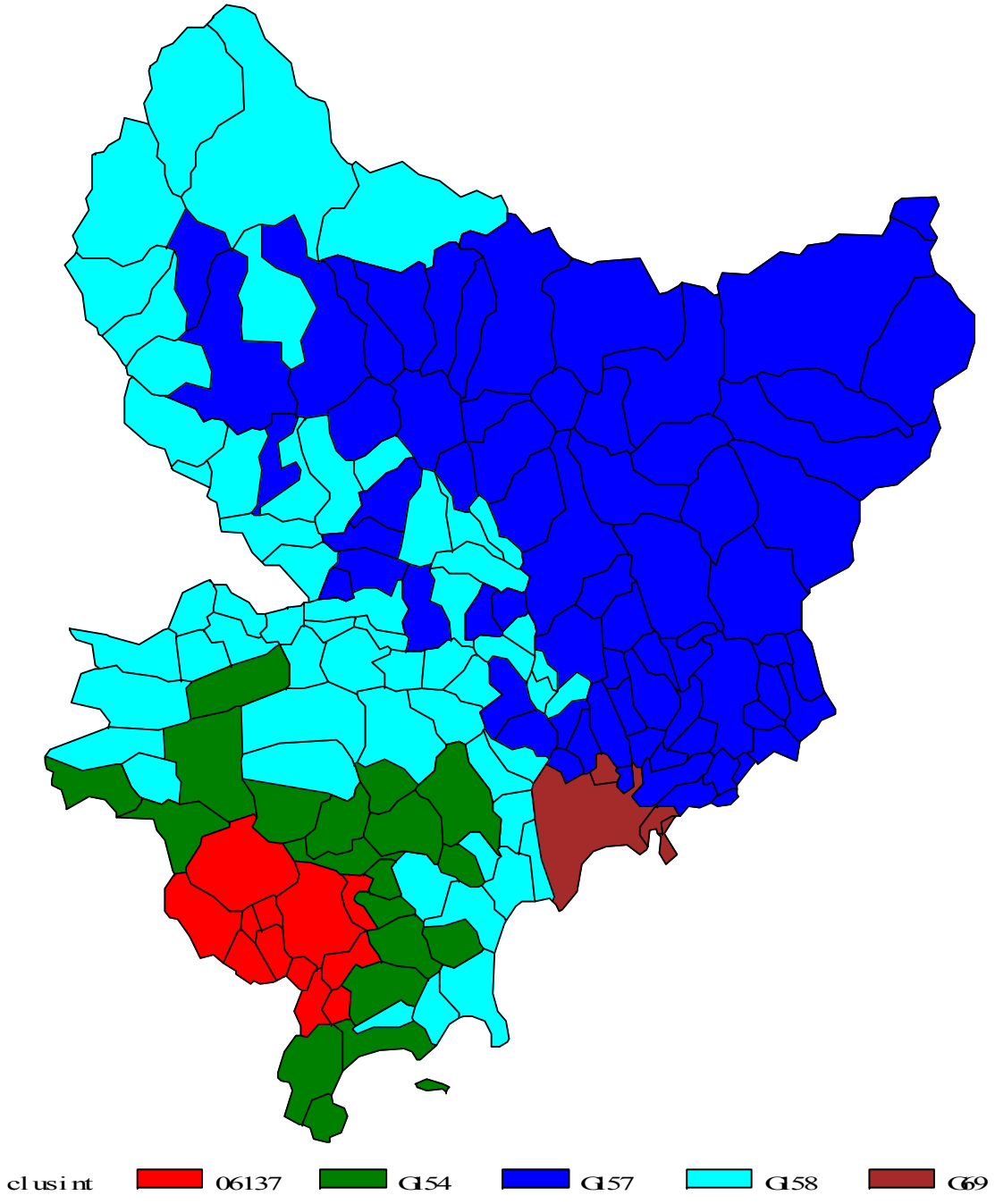
Phase 1 – Maximisation



Résultats de la première phase

Groupes	Taille	Age moyen	Variance intra
06137	0,1%	43,5	0
G154	26,2%	43,5	2,692
G157	8,2%	43,9	0,566
G158	22,0%	42,7	2,158
G69	30,0%	43,0	0,544
		43,1	**5,960**

Phase 2 – Maximisation



Résultats de la seconde phase

Groupes	Taille	Age moyen	Variance intra
06137	9,2%	40,4	0,370
G154	16,5%	43,8	2,216
G157	15,0%	42,7	1,376
G158	26,2%	43,8	0,897
G69	33,1%	43,2	0,309
		43,1	**5,168**

Extensions possibles

- Groupes *les plus hétérogènes possibles*
  - Variance intra-groupe à maximiser et non plus à minimiser
- Extensions à plusieurs variables
- Extensions à des distances non euclidiennes

5. Une extension de la méthode

- L'inertie d'un groupe *P* peut se réécrire :

$$I = \frac{1}{2\omega^2} \sum_{i,j \in P} \alpha_i \alpha_j (x_i - x_j)^2$$

- Les notations sont explicitées dans le papier

- Ce qui peut se réécrire :

$$I = \frac{1}{2\omega^2} \sum_{i,j \in P} \alpha_i \alpha_j d_{ij}^{\ 2} \quad (2)$$

Une nouvelle interprétation

- La formule (2) permet d'étendre la méthode à des distances non euclidiennes
- Mais ne permet pas d'utiliser les centres de gravité de chaque groupe (calculs plus complexes)

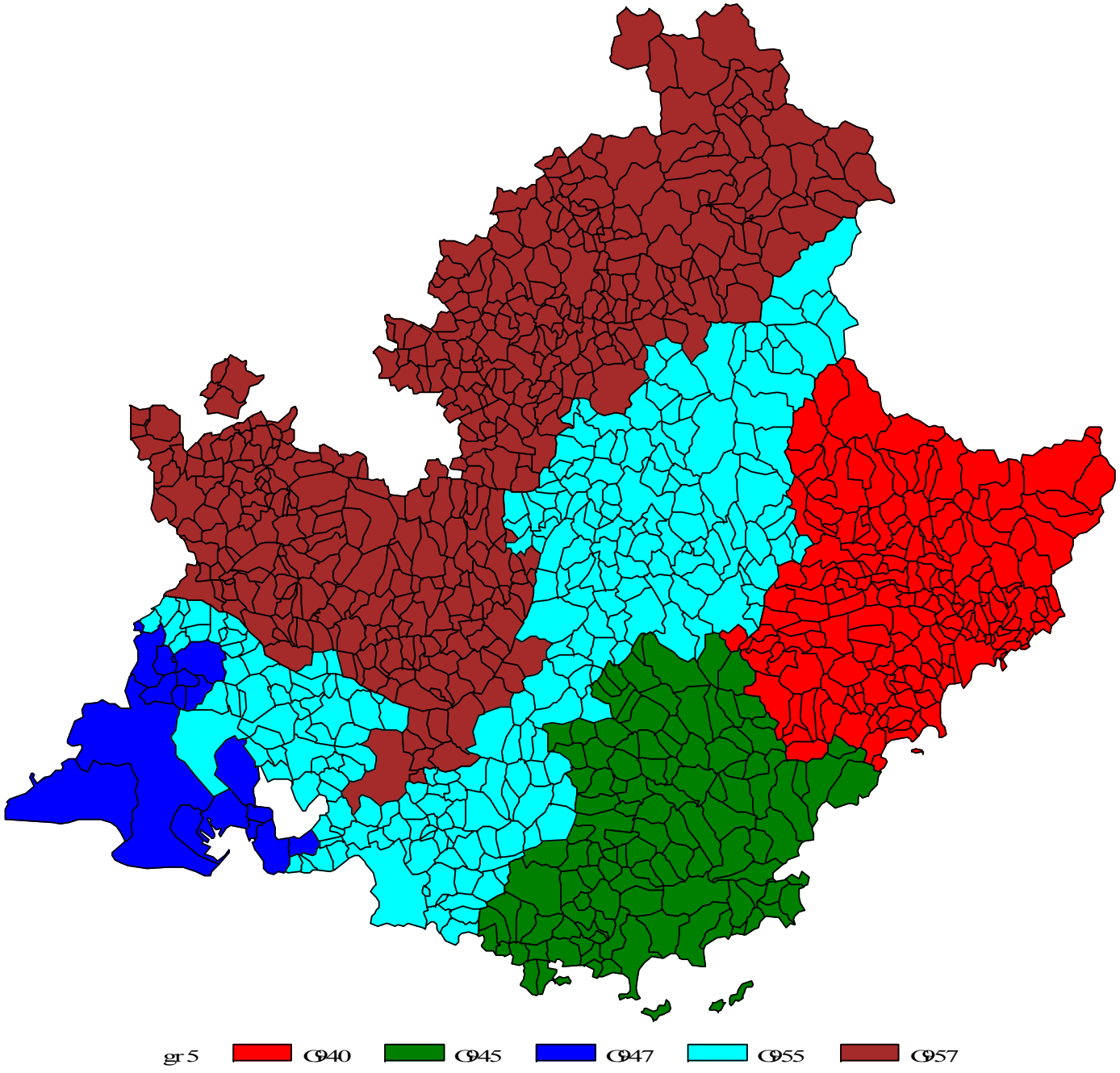
Navettes domicile-travail en PACA

- On cherche à créer des groupes connexes de communes qui maximisent en un certain sens la proportion de personnes résidant et travaillant dans la même zone.
  - Un calcul montre qu'il suffit de prendre

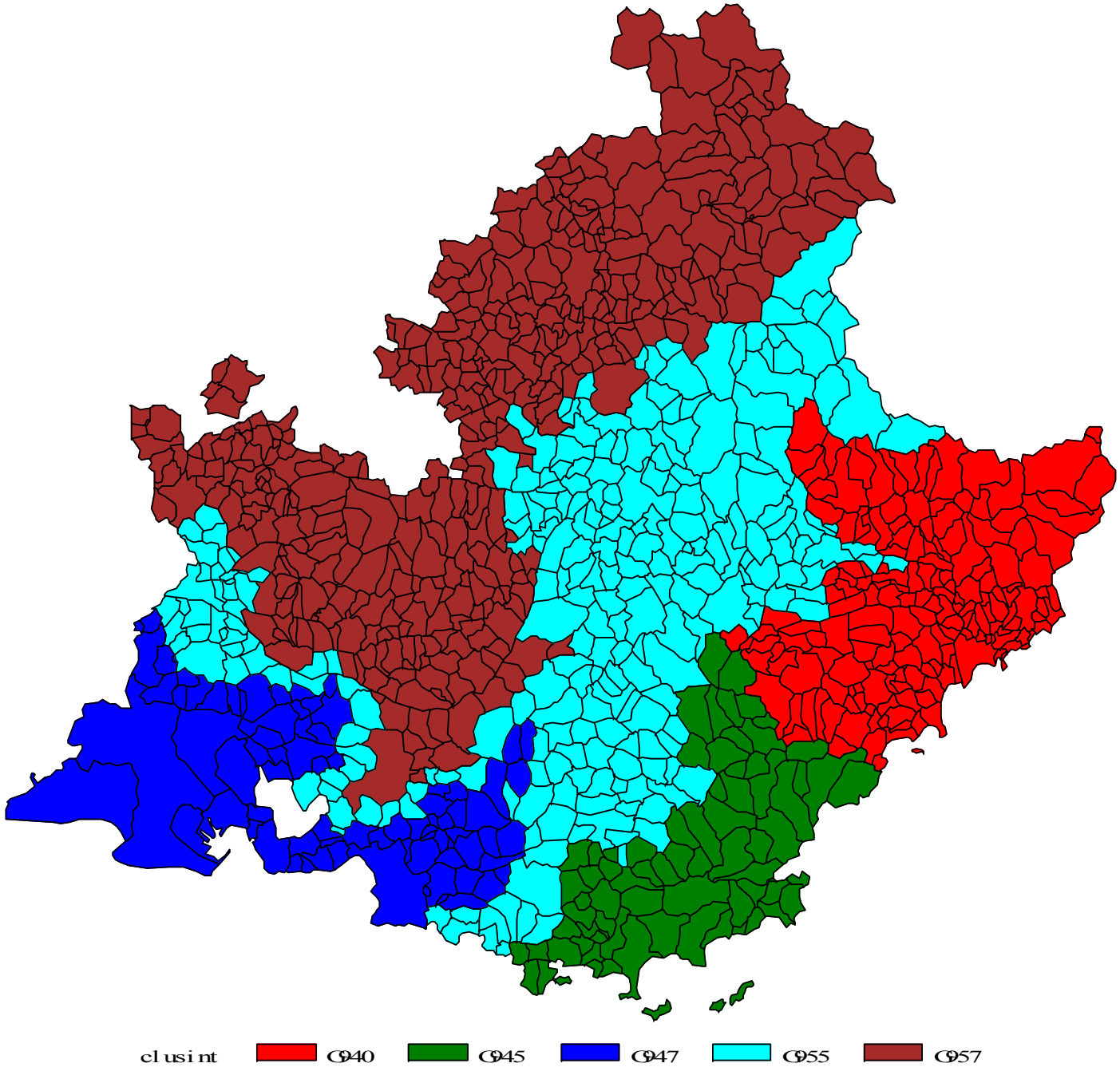
$$\begin{cases} \alpha_i = A_{i,\bullet} \\ d_{i,j} = \sqrt{\dfrac{A_{i,j}}{A_{i,\bullet} A_{j,\bullet}}} \end{cases}$$

  - Où A_{ij} représente la population résidant dans la commune i et travaillant dans la commune j)

Navettes résidence – travail avant optimisation



Navettes résidence–travail après optimisation



Les résultats

	Phase 1	Phase 2
%personnes travaillant et résidant dans la même zone	91,4%	88,2%
Taille minimale des zones	5,7% (G947)	17,1% (G955)
Taille maximale des zones	28,1%(G940)	28,1%(G940)

6. Conclusion

- Une macro SAS est disponible sur demande
- Il y actuellement des limitations sur le nombre d'unités statistiques à traiter => des améliorations du programme sont à l'étude
- D'autres applications doivent être testées et leurs résultats analysés :
  - Application à des données de flux (déplacements domicile –travail, comparaison lieu de naissance / lieu de résidence..)
  - Affectation d'un échantillon entre différents enquêteurs minimisant un critère de temps ou de distance de déplacement.

Merci de votre attention

Marc.christine@insee.fr
Michel.isnard@insee.fr